RESEARCH ARTICLE

RSP Science Hub

International Research Journal on Advanced Science Hub
2582-4376
Vol. 04, Issue 03 March
www.rspsciencehub.com

Check for updates

http://dx.doi.org/10.47392/irjash.2022.014

# Review on Intrusion Detection System (IDS) for Network Security using Machine Learning Algorithms

*Nusrath Unnisa A*[1], *Manjula Yerva*[2], *M Z Kurian*[3]

[1]*Department of Electronics and Communications Engineering, Sri Siddhartha Institute of Technology, Karnataka, India*
[2]*Assistant Professor, Department ofElectronics and Communications Engineering, Sri Siddhartha Institute of Technology, Karnataka, India*
[3]*Head of the Department, Department of Electronics and Communications Engineering, Sri Siddhartha Institute of Technology, Karnataka, India*

Emails: nussukhanum@gmail.com, manjulayerva@ssit.edu.in, mzkurianvc@yahoo.com

## Article History

*Received: 14 February 2022*
*Accepted: 25 March 2022*

***Keywords***:

*Denial-of-Service;*
*Intrusion detection system;*
*Machine learning algorithms;*
*Network;*
*User-to-Root;*
*Remote-to-Login*

## Abstract

*With the advancement in the artificial intelligence technologies and development of fifth generation networks, a network may face many hazards and challenges as the number of users are accessing the network simultaneously which makes the user to think of losing the confidentiality of the data and hence the network to be considered for security. Threats on the network can be classified in many ways and to detect such threats an Intrusion detection system (IDS) is the one which is mainly used. A network can be attacked in two ways as minor attack and major attack. Denial-of-Service (DoS) and Prob attacks belong to major kind and User-to-Root (U2R) and Remote-to-Login (R2L) goes to minor attack categories. The minor attacks are also called as rare attacks which are very injurious for a host and it is very difficult to recognize these attacks. This paper consists of a survey made on IDS and different algorithms used to implement these IDSs using machine learning.*

## 1. Introduction

In today's modern life, networks play a vital role which makes the security of internet as an interesting field of research. Many methods exist for network security such as firewalls, anti-malware software and Intrusion detection systems (IDSs), which helps the networks to protect against internal and external storming.

Along with this, in order to provide security of network by taking care of the software and hardware on the network an IDS is used. In 1980 the first kind of intrusion detection system (Pattawaro and Polprasert) came to existence and since then many IDS methods were developed. But the complications with available IDSs were they were unable to find the threats which were happening unknowingly, such as switching in the network environment. In order to find solution to these problems, human independent IDSs were developed which are built on the machine learning techniques. Machine learning belongs to the family of artificial intelligence which allows software applications to produce more accurate outcome without human intervention by training a machine how to learn (Anita, S, and Gupta). Based on the sample data which is also called as training data these algorithms build a model in order to make decisions without being explicitly programmed to do so.The objective of this paper is to figure out the concept of IDSs and to dis-

cuss distinct machine learning algorithms which can be used to implement IDSs.

## 2. Concept of IDS

An application software used to take care of a network or system for malicious activity is called as Intrusion detection systems. These IDSs are very crucial for a network which monitors the host and networks, generate alarm based on the behavior of computer systems and also for any suspicious activity.

Intrusion detection systems acts as heart of a network structure whose work is to keep on monitoring the network activities by regularly verifying the connection patterns and flow of packets through that network (Buczak and Guven). The advantage of these IDS is their ability to classify the network and packets based on the set of predefined parameters (Shah et al.).

Based on the recognition methods used, the IDSs can be basically classified as detection based methods and source based methods. Under detection based methods there are two categories such as signature based detection (also called as misuse detection) and anomaly detection. For the source based methods also there exists two classifications such as host based and network based methods (Masduki et al.)

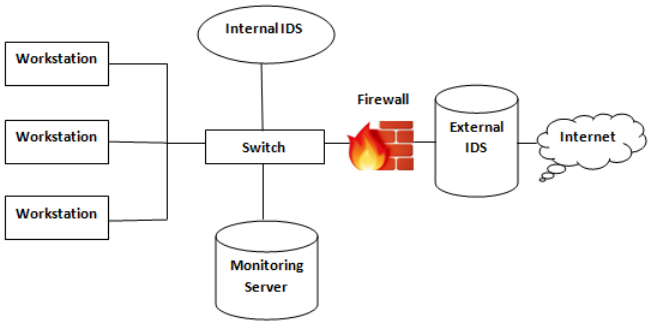


**FIGURE 1. An Intrusion Detection System Model**

### 2.1. Detection based methods

As previously mentioned, these methods falls into two categories such as signature based detection and anomaly detection.

#### 2.1.1. Signature based Detection

This method is generally known as signature based methods which is rule based (Umbarkar and Shukla). This kind of detection methods relies on signature database which works by comparing the sample of signature with those in the database. But the drawback of this method is to compose well organized signatures (Lansky et al.). This method gained popularity because it gives the reports of attack types with the reasons caused and also it provides low false alarm rate. Coming to the disadvantage, this method has elevated missed alarm rate and it is poor on detecting the unknown attacks and maintaining a vast signature data is also needed.

#### 2.1.2. Anomaly Detection

This kind of detection method is used to detect change in behavior. The working principle is to create a normal behavior profile and comparing the activity against that profile. If there is any deviation it triggers an alert. Therefore, it is mandatory in this detection method to create a normal profile (Khosravi-Farmad, Ramaki, and Bafghi). Benefit of this method is that it supports good generic and it is very efficient in finding unknown attacks. Shortcoming is that false alarm rate is more and it is inefficient to provide the reasons for irregularity. Distinction between signature based detection and anomaly detection is listed in Table1.

### 2.2. Source of data based methods

Under this category there exist two methods such as host based IDS and network based IDS.

#### 2.2.1. Host based IDSs

These are the entities which are based on software and installed on a host computer to inspect and take care of all the congestion happening on the system application files and operating systems. Benefit of this method is their ability to detect hazards by checking the congestion happening in the network before exchanging of data. The disadvantage is that only the host system is monitored i.e these IDSs has to be put on every host (P. Singh, S. P. Singh, and D. S. Singh).

#### 2.2.2. Network based IDSs

In contrast to host based IDSs, these IDSs are hardware based and need to be installed to monitor network congestion. Since these IDs are not dependent on OS, they can be installed in any OS environment (Hindy et al.). Benefit of this method is their ability to find types of protocol specifically and attacks on network. But the shortcoming is only the network segment through which traffic is moving

**TABLE 1. Differentiation among signature based detection and anomaly detection**

| | Signature based | Anomaly Detection |
|---|---|---|
| Performance Recognition | It has gaint missed alarm rate and false alarm rate is small | It has gaint false alarm rate and missed alarm rate is small |
| Proficiency Detection | It has elevated efficiency detection and it reduces with signature baseline scale. | Provide efficiency based on the complexity of the model |
| Domain Knowledge dependency | All findings are dependent on the domain knowledge | Provides low domain knowledge dependency because only the characteristics design hang on domain knowledge. |
| Explication | Since designs hang on the domain knowledge provides strong interpretative. | Provides weak interpretative because it gives only detected results. |
| Unspecified attack Recognition | Only known attacks are detected. | Both known and unknown attacks are detected. |

will be taken care. Differences between host-based and network-based IDSs are listed in table 2.

## 3. Machine Learning Algorithms used in IDS

Machine learning belongs to the family of artificial intelligence which uses software application to produce more accurate results without being explicitly programmed to do so. Based on historical data they produce new output values (Zhengbing, Zhitang, and Junqi). Since it is data driven method, knowing the type of data acts as the basic step. This section is to know about various machine learning algorithms used to implement IDSs. Support as SVM (support vector machine), KNN (K-nearest neighbor), ANN (artificial neural networks), LR (Logic regression), Naïve Bayes, decision tree, clustering and hybrid methods.

### *3.1. Artificial Neural Network (ANN)*

The design idea of ANN is based on the brain activities derived from the human behavior. These are the collection of connected nodes called as artificial neurons. An ANN consists of input layer, hidden layers and output layer, where all the nodes are fully connected. In order to get work done all these nodes need to be trained before implementation, but training these nodes is very time-consuming because of its tedious structure. Training of ANN models is generally done using back propagation algorithm which cannot be used in deep network training (Gu et al.).

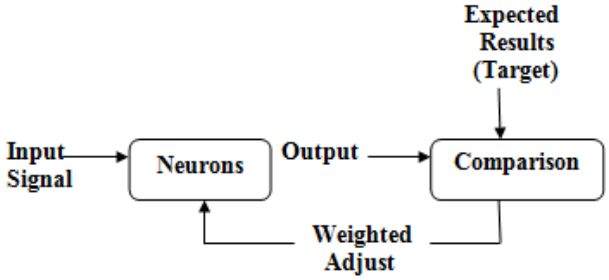


**FIGURE 2. Block diagram- ANN for pattern classification.**

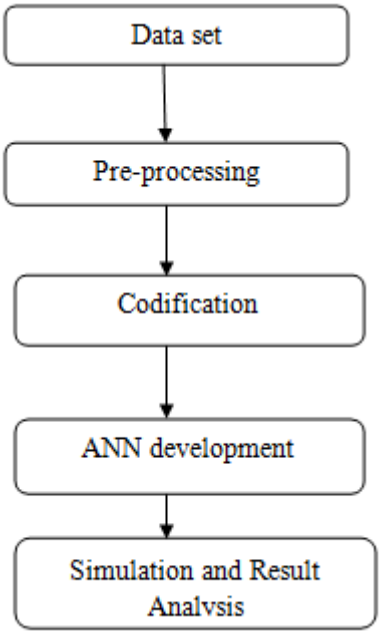


**FIGURE 3. Diagram of Activities in an ANN model**

**TABLE 2. Differentiation among host-based and network-based IDSs**

| | Host-Based IDS | Network-Based IDS |
|---|---|---|
| Data Source | Depends on operating system or application program records | Depends on Network congestion |
| Categorization | Since every host is to be equipped with IDs , difficult to categorize. | Easy to categorize since it depends on network models. |
| Efficiency of finding | Less efficient because it has to check number of records each time. | Since attacks are found in real time these are highly efficient. |
| Traceability of Intrusion | Intrusions are traced based on the system call paths. | Based on IP address and time stamps, time and position of intrusion is detected. |
| Disadvantage | Behavior of network cannot be analyzed. | Traffic only through the particular network segment is taken care. |

### *3.2. Support Vector Machine (SVM)*

The procedure of using SVM is to differentiate max margin hyperplane in the n-dimension feature space. Since the separation of hyperplane can be done even with small amount of support vectors these methods provide great results (Pressley). But near the hyperplane these methods are delicate to noise and also they are capable of solving linear problems.

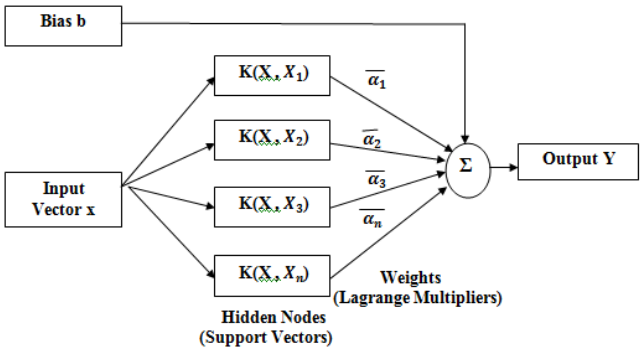


**FIGURE 4. Schematic diagram of SVM structure**

### *3.3. K-Nearest Neighbor (KNN)*

This is one of the basic method of classification which is non -parametric and very efficient in classification. The sample is said to have probability of belonging to the class if most of its neighbors belong to the same class (Tirumala, Sathu, and Sarrafzadeh). The performance of KNN is dependent on parameter K which is found by the user. According to the sample test taken, K training points are selected by considering the nearest distance to the test sample and hence the performance of KNN is greatly dependent on the parameter K. Based on the value of K complexity and fitting ability of model is decided.

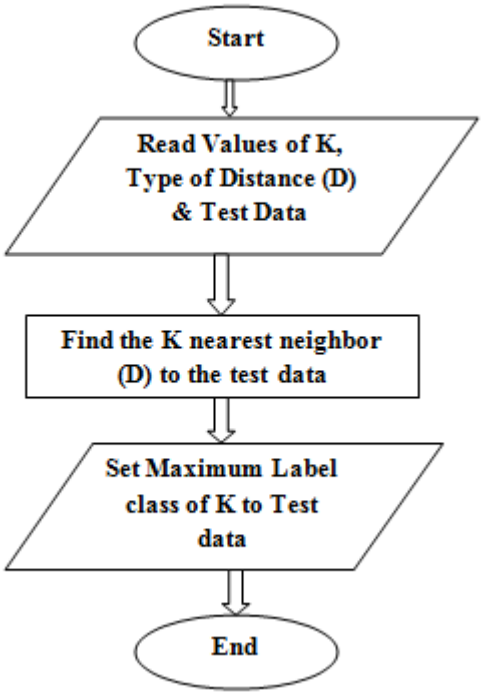


**FIGURE 5. KNN classification Algorithm**

### *3.4. Logistic Regression (LR)*

For solving the classification problem these models are used which works by computing the probability of different classes by considering the parametric logistic distribution (Dias et al.). These models are efficient due to their capacity of furnishing probabilities and to divide new data based on continuous and discrete datasets. Logistic distribution is calculated with the formula given below.

$$P(Y= k—x) = \frac{e^{x*w_k}}{1+\sum_k^{K-1} e^{x*w_k}} \quad .....1$$

where k = 1,2...K-1.

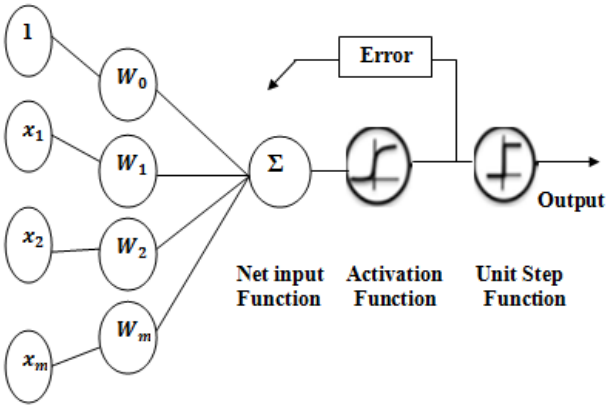


**FIGURE 6.** Logistic Regression model

### 3.5. Naive Bayes

Working principle of this method is conditional probability and attributes independence hypothesis (Shi et al.). These network acts as directed acrylic graphs in which every node acts as a discrete random variables of interest. The conditional probability table (CPT) is maintained, where each node carry the random variable state, which is used to specify the conditional probability of the domain variables with other connected variables. The formula for conditional probability is given below.

$P(X = x \mid Y = C_k) = \prod_{i=1}^{n} P(X^i = x^i \mid Y = C_k)$.....2

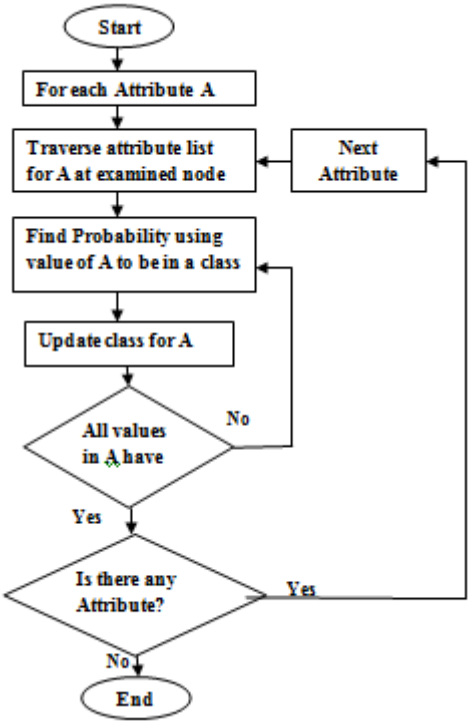


**FIGURE 7.** Naive Bayes classification

### 3.6. Decision tree

These algorithms are used in classification problems for learning and modelling a dataset. Classification of new data set is done based on what it has got from the previous dataset (Wasi et al.). By using decision tree algorithms , inappropriate and unnecessary features are excluded automatically. The learning steps involved in this model is, selecting the feature, generating tree and tree pruning. In order to train a decision tree model most appropriate features are selected individually and child nodes are generated from it.

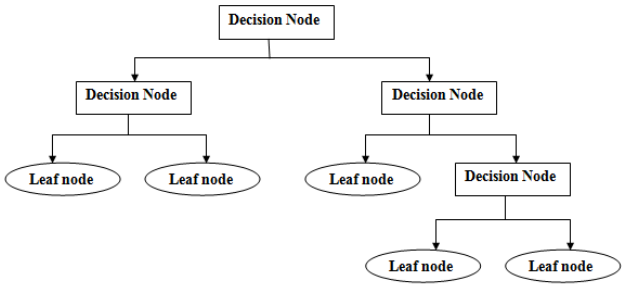


**FIGURE 8.** Decision Tree Hierarchy

### 3.7. Clustering

It is the task of segregating the data points into a number of groups, such that highly similar data are grouped into one cluster and less similar data are grouped into another cluster. Benefits of this algorithm are they don't need prior knowledge. It is mandatory to refer external information while detecting attacks using clustering algorithm. K-means is an example of clustering algorithm (Chen et al.).

**K-means**:It is s an exemplar of clustering algorithm which makes the use of Euclidean distance to compute centre of cluster and data. K refers to the number of clusters and mean stand for attribute mean. The idea behind this algorithm is to achieve less distance inside the same cluster and to have maximum distance between clusters. Advantages and disadvantages of various forward propagation models (shallow models) are listed in Table 3.

## 4. IDS Feature Selection

With respect to the machine learning classification two phases are involved such as classification phase and training phase. Dispersal of the characteristics is done in the course of training phase and

**TABLE 3. Advantages and disadvantages of various shallow models**

| Algorithms | Advantages | Disadvantages |
|---|---|---|
| **ANN** | Have good fitting ability and they are used to deal with non linear data. | *Training of this model is time consuming. *Liable to overfitting; |
| **SVM** | Even with small train set useful information is learnt. They have strong generation ability. | *Difficult to perform on big dataset using this algorithm. *For kernel related parameters they are very delicate. |
| **KNN** | *Easy to train *Best suited for non linear data. *For massive data also implemented *Strong against noise. | * Testing time is more. *Delicate to the K parameter. |
| **Naive Bayes** | *Incremental data learning can be done. *strong against noise. | Difficult to perform on attribute-related data. |
| **LR** | *Implementation is simple. *Training can be done very fast. | *Cannot perform on nonlinear data; *Liable to overfitting |
| **Decision tree** | *Selection of features can be done automatically. *Interpretation is very strong. | Correlation of data is ignored. |
| **K-means** | *Implementation is simple. *training of data can be done rapidly. | *Initialization is delicate. *Delicate to the parameter K |

learned features are put in as normal profile during the classification phase where any abnormality will be detected.

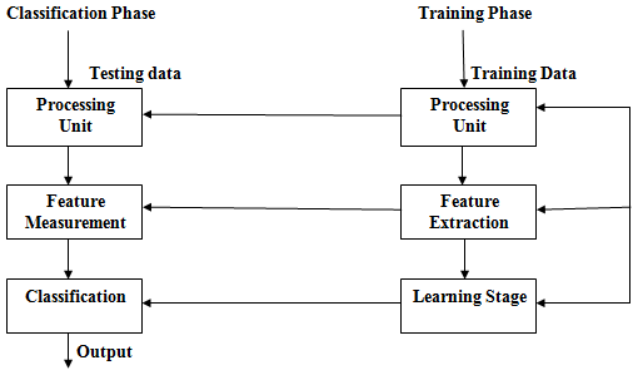


**FIGURE 9. Design of Machine learning classification process**

## 5. Conclusion

Classification of machine learning algorithms is focused in this survey and also machine leaning based IDSs are summarized which are implemented in the security of networks.

## ORCID iDs

Nusrath Unnisa A https://orcid.org/0000-0002-2452-9481

## References


Anita, Chordia, S, and S Gupta. "An effective model for anomaly IDS to improve the efficiency". *2015 International Conference on Green Computing and Internet of Things (ICGCIoT)* (2015): 190–194. 10.1109/ICGCIoT.2015.7380455.

Buczak, Anna L. and Erhan Guven. "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection". *IEEE Communications Surveys & Tutorials* 18.2 (2016): 1153–1176. 10 . 1109 / comst . 2015 . 2494502.

Chen, Mingzhe, et al. "Artificial Neural Networks-Based Machine Learning for Wireless Networks: A Tutorial". *IEEE Communications Surveys & Tutorials* 21.4 (2019): 3039–3071. 10.1109/comst.2019.2926625.

Dias, L P, et al. "Using artificial neural network in intrusion detection systems to computer networks". *9th Computer Science and Electronic Engineering (CEEC)* (2017): 145–150. 10.1109/CEEC.2017.8101615.

Gu, Bin, et al. "Kernel Path for $\nu$-Support Vector Classification". *IEEE Transactions on Neural Networks and Learning Systems* (2021): 1–12. 10.1109/tnnls.2021.3097248.

Hindy, Hanan, et al. "A Taxonomy of Network Threats and the Effect of Current Datasets on Intrusion Detection Systems". *IEEE Access* 8 (2020): 104650–104675. 10.1109/access.2020.3000179.

Khosravi-Farmad, Masoud, Ali Ahmadian Ramaki, and Abbas Ghaemi Bafghi. "Risk-based intrusion response management in IDS using Bayesian decision networks". *2015 5th International Conference on Computer and Knowledge Engineering (ICCKE)* (2015): 307–312. 10.1109/ICCKE.2015.7365847.

Lansky, Jan, et al. "Deep Learning-Based Intrusion Detection Systems: A Systematic Review". *IEEE Access* 9 (2021): 101574–101599. 10.1109/access.2021.3097247.

Masduki, Bisyron Wahyudi, et al. "Study on implementation of machine learning methods combination for improving attacks detection accuracy on Intrusion Detection System (IDS)". *2015 International Conference on Quality in Research (QiR)* (2015): 56–64. 10.1109/QiR.2015.7374895.

Pattawaro, Apichit and Chantri Polprasert. "Anomaly-Based Network Intrusion Detection System through Feature Selection and Hybrid Machine Learning Technique". *2018 16th International Conference on ICT and Knowledge Engineering (ICT&KE)* (2018): 1–6. 10.1109/ICTKE.2018.8612331.

Pressley, T. "A new paradigm for intrusion detection systems". *Proceedings. Eleventh International Conference on Computer Communications and Networks* (2002): 390–390. 10.1109/ICCCN.2002.1206523.

Shah, Ajay, et al. "Building Multiclass Classification Baselines for Anomaly-based Network Intrusion Detection Systems". *IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)* (2020): 759–760. 10 . 1109 / DSAA49011.2020.00102.

Shi, Kansheng, et al. "An improved KNN text classification algorithm based on density". *2011 IEEE International Conference on Cloud Computing and Intelligence Systems* (2011): 113–117. 10 . 1109/CCIS.2011.6045043.

Singh, Preeti, S P Singh, and D S Singh. "AN INTRODUCTION AND REVIEW ON MACHINE LEARNING APPLICATIONS IN MEDICINE AND HEALTHCARE". *2019 IEEE Conference on Information and Communication Technology* (2019): 1–6. 10 . 1109 / CICT48419 . 2019.9066250.

Tirumala, Sreenivas Sremath, Hira Sathu, and Abdolhossein Sarrafzadeh. "Free and open source intrusion detection systems: A study". *2015 International Conference on Machine Learning and Cybernetics (ICMLC)* (2015): 205–210. 10.1109/ICMLC.2015.7340923.

Umbarkar, Swapnil and Sanyam Shukla. "Analysis of Heuristic based Feature Reduction method in Intrusion Detection System". *2018 5th International Conference on Signal Processing and Integrated Networks (SPIN)* (2018): 717–720. 10 . 1109/SPIN.2018.8474283.

Wasi, Sarwar, et al. "Intrusion Detection Using Deep Learning and Statistical Data Analysis". *2019 4th International Conference on Emerging Trends in Engineering, Sciences and Technology (ICEEST)* (2019): 1–5. 10 . 1109 / ICEEST48626 . 2019 . 8981688.

Zhengbing, Hu, Li Zhitang, and Wu Junqi. "A Novel Network Intrusion Detection System (NIDS) Based on Signatures Search of Data Mining". *First International Workshop on Knowledge Discovery and Data Mining (WKDD 2008)* (2008): 10–16. 10.1109/WKDD.2008.48.



**Embargo period:** The article has no embargo period.

**To cite this Article:** Unnisa A, Nusrath, Manjula Yerva, and M Z Kurian. "**Review on Intrusion Detection System (IDS) for Network Security using Machine Learning Algorithms**." International Research Journal on Advanced Science Hub 04.03 March (2022): 67–74. http://dx.doi.org/10.47392/irjash.2022.014